# 日内中心矩因子

## 因子逻辑

### 概述

数学上，“矩”是一组点组成的模型的特定的数量测度。均值是一阶矩，方差是二阶矩。也是波动率因子的核心构造函数。偏度则对应三阶矩，峰度为四阶矩。用以描述数据样本的分布信息。

一阶矩到四阶矩，目前学术上都存在其相应的经济意义的。一阶矩代表收益，二阶矩代表风险，三阶矩代表分布的偏态，四阶矩代表分布的聚集程度。

一阶矩因子是经典的动量反转因子，在前面的研究中已经覆盖。

传统的波动率因子的以标准差为核心，去求诸如收益率，成交量等量价信息的标准差，去获取低波动股票的超额收益，由股票的低波动属性为核心构造的低波动指标是一个比较稳定有效的技术指标。

而前面日内反转因子的研究看出，高波动能够放大反转因子的收益，而本篇报告所研究的低波动指标恰好跟这种性质是相反的。在近年来传统反转因子出现回撤的时候，低波动因子也是一种比较有效的补充。

而关于偏度和峰度的研究近年来也有不少相关的研究报告，因此此处将其一并纳入本报告中。

### 研究方法

此前的研究发现，日内不同时间段具有不同的交易行为和交易结构，所以传统的日频反转因子无法反应日内交易行为结构的差异，所以考虑引入HFBar数据，去构造这种差异，反应日内不同的交易结构所蕴含的信息。

由于此前的研究报告中，包括日内流动性因子，日内反转因子的深化研究中，我们主要通过如下两个步骤来寻找有效的日内高频因子，再研究波动率因子时我们沿用这一种研究模式：

（1）寻找日频意义下比较有效的低波动因子构造模式，使之表现比当前FactorPool中因子有改善，或者是与当前FactorPool中的因子收益相关性较低。

（2）将日频下初步有效的因子高频化成日内波动因子，验证其的有效性。

## 日频下有效的因子模式的寻找

这一部分主要是通过现有的研报的信息，找寻发掘一些比较有效的日频波动率因子，我们不要求这部分因子要完全超越FactorPool中现有Volatility类因子的表现，只是希望能够从中找到一些能达到Volatility同类因子中相对优秀的水平，或者现有FactorPool中未能覆盖的alpha源，为我们进行日后高频因子的寻找提供比较优秀的素材。

《国泰君安数量化专题之一百零八：基于不同域研究的多因子选股体系》

《国泰君安数量化专题之九十三：基于短周期价量特征的多因子选股体系》

经过测试后，此处整理选取一些比较有效而且稳定的反转因子改进模式，列出如下

### 低波动因子Std

该因子是低波动因子最基础的定义，其构造逻辑是

-ts\_Stdev(Data,MAWin)

因子的基础定义来自如下研报《华泰单因子系列之五单因子测试之波动率因子》

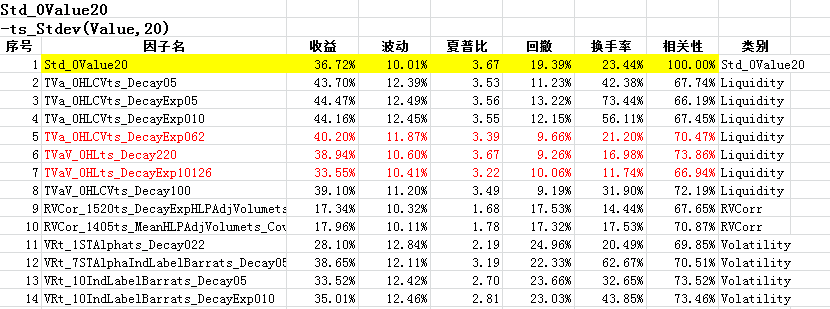
其中Data可以是TotalRet，Value，Volume，Turnover

我们进行了如下测试，我们测试如下形式的基础形式的低波动因子

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| -ts\_Stdev(Volume,20) | 24.02% | 12.44% | 1.93 | 29.26% | 17.43% |
| -ts\_Stdev(Value,20) | 36.72% | 10.01% | 3.67 | 19.39% | 23.44% |
| -ts\_Stdev(TotalRet,20) | 3.59% | 17.69% | 0.20 | 55.87% | 31.32% |
| -ts\_Stdev(TO,20) | 24.58% | 14.74% | 1.67 | 24.60% | 18.06% |

从上述结果可知，可以看出，对于成交额和换手率等信息来说，低波动因子是一类比较有效的因子。从其收益换手比来看，是一类不错的因子。

特别地，ts\_Stdev（Value,20）能够接近FactorPool中同类因子的比较优秀的水平，但是表现稍差于TVaV类因子（标红部分）。但是可以作为研究的基础因子。



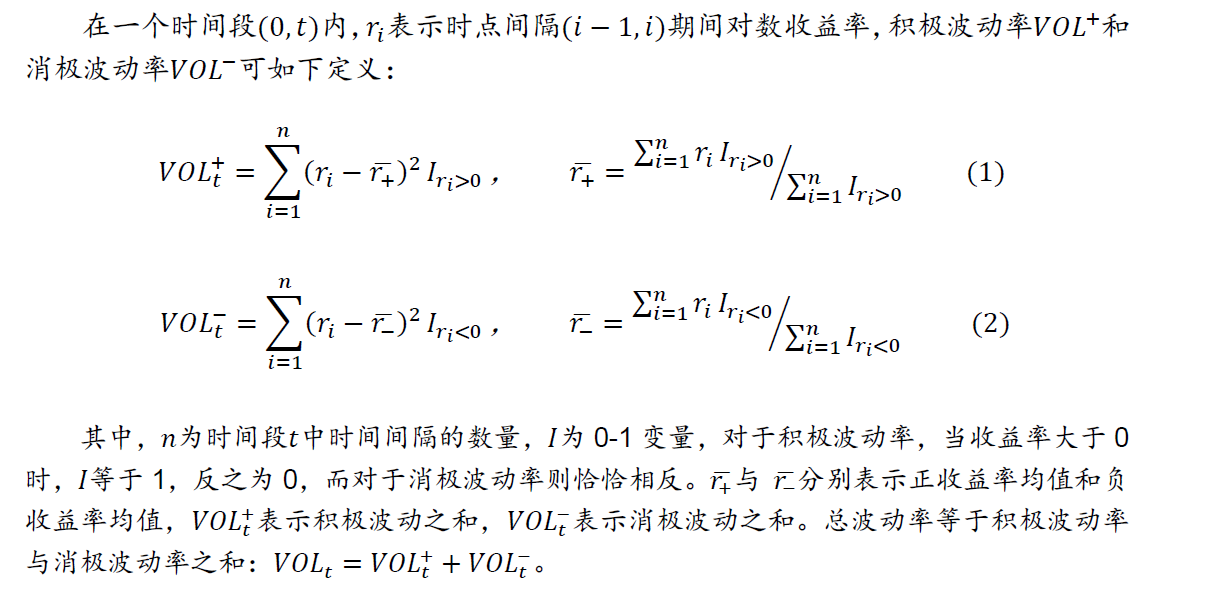
### 上下行波动率

传统的波动率因子在构造的过程中，对收益方向是不作区分的，但是在实际市场中，投资者特别是机构投资者则更多表现为风险厌恶型，即相较于盈利，对亏损更为敏感，但是传统的波动率因子并不能反映这一点，所以为了刻画此种性质，学术上构造了上下行波动率，分别刻画市场上行的市场波动，和市场下行的市场波动。

具体改进的基础思想来自以下两篇的研报

《东方证券-金融工程研究技术类新Alpha因子的批量测试 20170217》

《西南证券-基于方向波动率的选股因子研究20171221》



我们构造上下行波动率之差，进行如下测试

|  |  |
| --- | --- |
| DirStdRet\_1TotalRetTotalRet15 | Power(ts\_Stdev(IfThen((TotalRet),TotalRet,NaN),5),1)-Power(ts\_Stdev(IfThen((TotalRet),NaN,TotalRet),5),1) |
| LFVol\_1TotalRet20 | -ts\_Stdev(TotalRet,20) |

测试结果如下

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| DirStdRet\_1TotalRetTotalRet15 | -33.74% | 10.31% | -3.27 | 339.01% | 139.64% |
| LFVol\_1TotalRet20 | 3.59% | 17.69% | 0.20 | 55.87% | 31.32% |

可以看出，原本并不是alpha因子的收益率的波动率在拆解为上下行波动率之差时，性质出现了显著的变化，由回测可以看出，能够贡献稳定的负收益，即下行波动与上行波动之差是一个稳定的alpha因子。

所以我们可以看见对波动率的这种分拆处理能够有效的挖掘原本不被发现的信息。

但是值得注意的是，这种处理目前只针对收益率有效，对成交量，换手等信息效果并不明显。

### 变异系数类（Coefficient of Variation）

此类因子（以下简称CV类因子）是对于波动率因子的另外一种改进，其目的是消除数据样本本身量纲的影响。

我们先看一个例子，先前的测试显示

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| -ts\_Stdev(Volume,20) | 24.02% | 12.44% | 1.93 | 29.26% | 17.43% |
| -ts\_Stdev(Value,20) | 36.72% | 10.01% | 3.67 | 19.39% | 23.44% |

基于成交额的波动率因子相较于基于成交量的波动率因子更为有效。而且两者数量级上存在这样一个关系式Value = Volume\*VWAP，即差了一个价格的量纲级

我们进行了以下测试，

LFVol\_1Volume20 ： -ts\_Stdev(Volume,20)

HLDU\_1Volume2020： -ts\_Mean((High-Low),20).\*ts\_Stdev(Volume,20)

LFVol\_1Value20：-ts\_Stdev(Value,20)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| LFVol\_1Volume20 | 24.02% | 12.44% | 1.93 | 29.26% | 17.43% |
| HLDU\_1Volume2020 | 34.36% | 10.87% | 3.16 | 17.21% | 23.60% |
| LFVol\_1Value20 | 36.72% | 10.01% | 3.67 | 19.39% | 23.44% |

可以看出，在-ts\_Stdev(Volume,20)乘以ts\_Mean((High-Low),20)，收益有明显提升，且其表现接近-ts\_Stdev(Value,20)，相关性达到0.96。

因此数据本身的量纲会对因子表现产生影响，所以我们在低波动因子基础上，引入变异系数这一个衍生指标。消除数据样本量纲的影响。

所以我们测试如下，测试变异系数类因子的表现

因子定义依次如下所示

LFVol\_4Volume20：-ts\_Stdev(Volume,20)./ts\_Mean(Volume,20)

LFVol\_4Value20： -ts\_Stdev(Value,20)./ts\_Mean(Value,20)

LFVol\_4TotalRet20：-ts\_Stdev(TotalRet,20)./ts\_Mean(TotalRet,20)

LFVol\_4TO20：-ts\_Stdev(TO,20)./ts\_Mean(TO,20)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| LFVol\_4Volume20 | 21.25% | 9.32% | 2.28 | 18.93% | 41.07% |
| LFVol\_4Value20 | 22.37% | 9.25% | 2.42 | 16.63% | 41.53% |
| LFVol\_4TotalRet20 | -14.96% | 7.60% | -1.97 | 153.64% | 144.76% |
| LFVol\_4TO20 | 21.00% | 9.35% | 2.25 | 19.09% | 41.07% |

可以看出，上述几种CV类因子，可以看出，除了利用TotalRet构造的变异系数以外，其他的综合表现还是一个可以的alpha因子，而且，去除量纲之后，他们的表现也比较相近，说明去除指标量纲影响，他们的alpha可能是同源的。

而且，经过测算，这样操作之后，因子的相关性与原低波动因子的相关性并不高，以-ts\_Stdev(Volume,20)./ts\_Mean(Volume,20)和-ts\_Stdev(Volume,20)两个因子为例，他们的相关性仅为0.26，所以可以作为一种有效的模式予以保留。

### 相对波动率

相对波动率因子的构造逻辑是个股短周期的波动水平与长周期的波动水平的相对位置关系的一个描述，因子的核心公式描述如下

-1\*Divide（ts\_Stdev(Data,SWin)，ts\_Stdev(Data,LWin)）

其中SWin为短平滑周期，LWin为长平滑周期

该因子认为，当短周期波动处于长周期波动的一个相对低位的时候，认为因子接下来将会表现更优，本质上仍然是一个低波动因子，只是相对于1中的绝对的波动因子来说，加入了长周期的一个波动性进行确认。

此种因子的改进思路来自于《国君数量化专题一零八—基于不同域研究的多因子选股体系》

我们进行了以下的测试：

因子定义依次如下：

|  |  |
| --- | --- |
| LSStd\_0Value20120 | -ts\_Stdev(Value,20)./ts\_Stdev(Value,120) |
| LSStd\_0TO20120 | -ts\_Stdev(TO,20)./ts\_Stdev(TO,120) |
| LSStd\_0TotalRet20120 | -ts\_Stdev(TotalRet,20)./ts\_Stdev(TotalRet,120) |
| LSStd\_0IdioRet20120 | -ts\_Stdev(IdioRet,20)./ts\_Stdev(IdioRet,120) |

测试结果如下

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| LSStd\_0Value20120 | 25.21% | 10.32% | 2.44 | 11.53% | 27.42% |
| LSStd\_0TO20120 | 21.59% | 9.06% | 2.38 | 10.63% | 27.78% |
| LSStd\_0TotalRet20120 | -2.60% | 11.32% | -0.23 | 66.06% | 38.74% |
| LSStd\_0IdioRet20120 | 21.48% | 9.49% | 2.26 | 12.81% | 32.51% |

可以看出，除了TotalRet以外，虽然收益波动并不突出，但其他的相对波动因子仍然是一种比较有效稳健Alpha因子模式。

### 高阶矩因子

波动率数学定义上反映的是数据样本离均值的离差平方和，是统计意义上的二阶矩，经济意义上反应的对资产风险大小的刻画。

而偏度和峰度是两个目前被赋予了经济意义的高阶矩因子。偏度反应的是数据分布的厚尾情况，而峰度是反应数据分布的聚集度情况。分别对应统计意义上的三阶矩和四阶矩。

因子的构造逻辑是：

-ts\_Skewness(Data,MAWin)

-ts\_Kurtosis(Data, MAWin)

以此为依据，我们构造了日频下的高阶矩指标，进行了如下测试

偏度因子定义依次如下

|  |  |
| --- | --- |
| LFVol\_2Volume20 | -ts\_Skewness(Volume,20) |
| LFVol\_2Value20 | -ts\_Skewness(Value,20) |
| LFVol\_2TotalRet20 | -ts\_Skewness(TotalRet,20) |
| LFVol\_2TO20 | -ts\_Skewness(TO,20) |

测试结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| LFVol\_2Volume20 | 7.89% | 7.09% | 1.11 | 17.14% | 59.24% |
| LFVol\_2Value20 | 9.95% | 6.93% | 1.44 | 13.87% | 60.29% |
| LFVol\_2TotalRet20 | 4.28% | 8.84% | 0.48 | 40.47% | 52.32% |
| LFVol\_2TO20 | 7.91% | 7.07% | 1.12 | 17.90% | 59.11% |

峰度因子的定义如下所示

|  |  |
| --- | --- |
| LFVol\_3Volume20 | -ts\_Kurtosis(Volume,20) |
| LFVol\_3Value20 | -ts\_Kurtosis(Value,20) |
| LFVol\_3TotalRet20 | -ts\_Kurtosis(TotalRet,20) |
| LFVol\_3TO20 | -ts\_Kurtosis(TO,20) |

测试结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| LFVol\_3Volume20 | 2.72% | 6.40% | 0.43 | 23.50% | 64.30% |
| LFVol\_3Value20 | 5.70% | 6.29% | 0.91 | 20.31% | 65.22% |
| LFVol\_3TotalRet20 | 8.95% | 10.29% | 0.87 | 31.94% | 59.19% |
| LFVol\_3TO20 | 2.86% | 6.34% | 0.45 | 22.71% | 64.14% |

从上述回测结果可以看出，日频形式的高阶矩因子表现非常一般，但是目前FactorPool中对高阶矩因子并没有覆盖，另外之所以保留是因为日内数据生成日内已实现高阶矩因子，效果有比较明显的改善。因此在此处先进行一个铺垫

## 因子高频化处理处理

在第二节中，我们找寻几种改进低波动指标的因子，下面我们对上述指标进行高频化处理，与反转因子不同，波动因子的高频化处理方法比较丰富，学术上也有不少研究。

以下我们就不同的日内化处理进行测试，并将测试中比较有意义的结果整理成小节，列出。其他表现平庸的结果我们再次不再赘述。

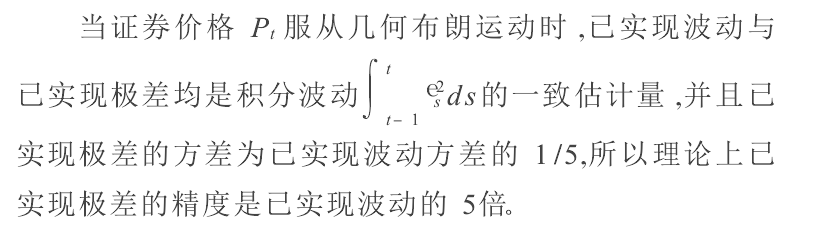
### 已实现测量

已实现测量是一种基于日内高频数据对日内波动的非参数估计方法。其核心逻辑将一段时间的日内信息对日内波动的一种无模型非参数估计。即已实现波动率

在后续的研究中，不少学者将已实现波动推广，得到已实现偏度和已实现峰度，构成了基于日内高频数据的已实现中心矩指标。

比较常见的已实现波动率的估计方法有两种，一种是由Anderson和Bollerslev（2001）提出的基于日内收益率的平方和的已实现波动率指标（Realized Volatility），另一种是Christensen（2005）在极差理论上给出的新估计量已实现极差波动率（Realized Range Volatility）

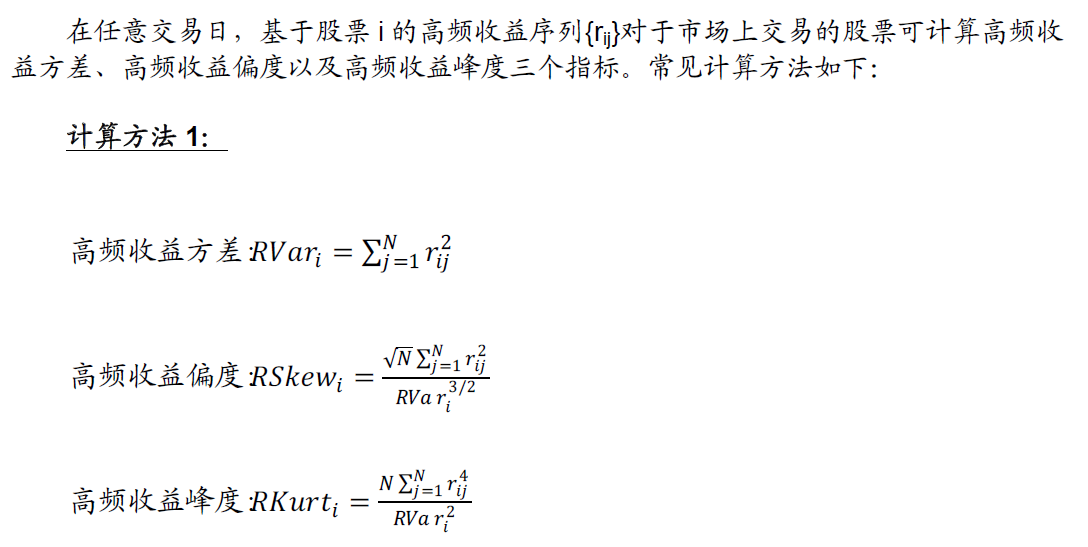
且学术均可以证明如下结论，这两种日内波动都是对波动率的无偏估计



我们将在接下来的篇幅测试以上两种不同的日内波动的估计方法和不同取样频率下构造下的因子表现。

**已实现波动（RV）**

基于日内收益率的平方和的已实现波动率指标的定义如下所示，其中可以将该方法推广到已实现偏度和已实现峰度。具体定义如下



基于1min收益数据，我们进行如下测试

因子定义如下：

|  |  |
| --- | --- |
| RVol\_75 | -ts\_Mean(IPRetp2,5) |
| Rskew\_15 | -ts\_Mean(Sqrt(240)\*IPRetp3./Power(IPRetp2,3/4),5) |
| Kurt\_15 | -ts\_Mean(Sqrt(240)\*IPRetp4./IPRetp2,5) |
| LFVol\_0TotalRet5 | -ts\_Stdev(TotalRet,5) |
| LFVol\_2TotalRet5 | -ts\_Skewness(TotalRet,5) |
| LFVol\_3TotalRet5 | -ts\_Kurtosis(TotalRet,5) |

其中IPRetp2日内1分钟收益率的平方和，类似的IPRetp3为立方和，IPRetp4为四次方和。

前三个因子是经过5天平滑处理的已实现波动，已实现偏度，已实现峰度因子，后面三个是日频下的波动，偏度和偏度因子。测试结果如下所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| RVol\_75 | 10.67% | 16.31% | 0.65 | 38.40% | 45.38% |
| Rskew\_15 | 34.58% | 8.71% | 3.97 | 10.53% | 98.15% |
| Kurt\_15 | 17.79% | 12.95% | 1.37 | 27.38% | 59.01% |
| LFVol\_0TotalRet5 | -12.82% | 14.85% | -0.86 | 172.32% | 98.40% |
| LFVol\_2TotalRet5 | 5.91% | 8.27% | 0.71 | 13.67% | 163.47% |
| LFVol\_3TotalRet5 | 1.52% | 6.87% | 0.22 | 23.70% | 186.91% |

可以看出高频化后的因子日内已实现因子效果相对于低频有明显提升，说明这种基于日内收益率的幂次加和是一种有效的日内化处理。

至于数据频率上面，我们测试了从1min到30min下的频率，

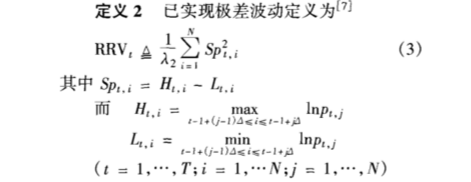
以已实现波动率因子为例，测试如下

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| ts\_Mean(IPRV1min,20) | -8.34% | 15.52% | -0.54 | 105.75% | 17.55% |
| ts\_Mean(IPRV5min,20) | -8.16% | 15.08% | -0.54 | 103.17% | 19.49% |
| ts\_Mean(IPRV10min,20) | -7.21% | 14.51% | -0.50 | 93.39% | 21.37% |
| ts\_Mean(IPRV15min,20) | -6.74% | 14.03% | -0.48 | 89.35% | 22.78% |
| ts\_Mean(IPRV30min,20) | -4.31% | 12.65% | -0.34 | 67.87% | 25.56% |

结论发现，频率采用1min先效果最好。（此处公式差一个符号，将收益和Sharp乘以一个-1理解即可）

**已实现极差（RRV）**

基于区间极差对日内波动进行估计的方法定义如下：



即将一个时间段分割为N个区间，每个区间的(High-Low)的平方和,除以常数4ln2即为该时间段的已实现极差的波动估计。

对于分割区间N的选取，我们进行如下测试

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| ts\_Mean(IPHLV1min,20) | 1.23% | 11.97% | 0.10 | 36.33% | 18.75% |
| ts\_Mean(IPHLV5min,20) | -6.57% | 14.52% | -0.45 | 77.38% | 19.08% |
| ts\_Mean(IPHLV10min,20) | -7.40% | 14.86% | -0.50 | 84.64% | 19.17% |
| ts\_Mean(IPHLV15min,20) | -7.68% | 14.96% | -0.51 | 87.51% | 19.29% |
| ts\_Mean(IPHLV30min,20) | -7.45% | 15.11% | -0.49 | 85.71% | 19.64% |

可以看出，10到15min之间的分割频率下是一个比较理想的选择。（此处公式同样差一个负号）

再进一步测试中，我们通过比较发现几类不同的低波动因子，使用已实现波动率构造波动率因子要稍微优于用已实现极差构造波动率因子。

但是并不绝对，例如以下因子，利用已实现极差优于已实现波动。

测试结果如下，IPHLV代表已实现极差构造的，IPRV代表利用已实现波动构造的

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| ts\_Stdev(IPHLV5min,20) | -24.85% | 14.76% | -1.68 | 252.27% | 24.47% |
| ts\_Stdev(IPHLV10min,20) | -24.40% | 14.95% | -1.63 | 247.17% | 24.51% |
| ts\_Stdev(IPHLV15min,20) | -24.38% | 14.99% | -1.63 | 247.66% | 24.55% |
| ts\_Stdev(IPRV1min,20) | -22.43% | 13.75% | -1.63 | 229.72% | 24.19% |
| ts\_Stdev(IPRV5min,20) | -18.97% | 13.13% | -1.45 | 194.90% | 24.74% |
| ts\_Stdev(IPRV10min,20) | -15.59% | 12.49% | -1.25 | 160.96% | 25.24% |
| ts\_Stdev(IPRV15min,20) | -14.02% | 11.97% | -1.17 | 145.43% | 25.51% |

**扩展一**

上文可以看到，采用日内收益率的幂次加和构造因子相较于原低频因子有比较大的改善，但是仍然有待改进的空间，主要问题在于：

1. 换手率较高
2. 与反转因子相关性较高

为了改善这两个问题，我们回到低波动因子的核心逻辑以及已实现波动率的构造核心

A．-ts\_Stdev(Data,MAWin)

B．

本质上是B对A的一种高频渐近，r是Close的分钟变化率，以此我们构造了基于分钟收益率（分钟收盘价的变化率）的已实现波动，已实现偏度，和已实现峰度

类似地，我们可以将这一概念从价扩展至量，构造基于成交量的变化率，成交额的变化率和换手率的变化率构造已实现指标。

以已实现偏度因子为例，

因子构造具体如下：

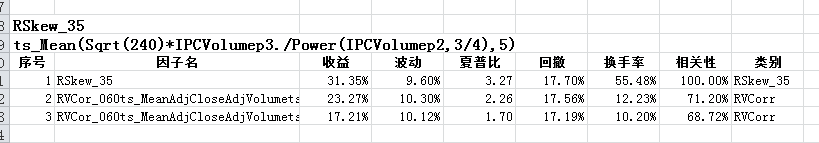
|  |  |
| --- | --- |
| RSkewVolume | ts\_Mean(Sqrt(240)\*IPCVolumep3./Power(IPCVolumep2,3/4),5) |
| RskewValue | ts\_Mean(Sqrt(240)\*IPCValuep3./Power(IPCValuep2,3/4),5) |
| RskewTO | ts\_Mean(Sqrt(240)\*IPCTOp3./Power(IPCTOp2,3/4),5) |
| RskewRet | -ts\_Mean(Sqrt(240)\*IPRetp3./Power(IPRetp2,3/4),5) |

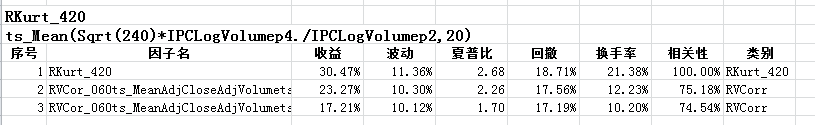
测试结果如下

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| RSkewVolume | 31.35% | 9.60% | 3.27 | 17.70% | 55.48% |
| RskewValue | 31.30% | 9.60% | 3.26 | 17.70% | 55.49% |
| RskewTO | 27.39% | 9.64% | 2.84 | 18.49% | 55.31% |
| RskewRet | 34.58% | 8.71% | 3.97 | 10.53% | 98.15% |

可以看出，这种改进下，因子特性在此处出现了比较有意思的事情，因子的相较于基于收益率构造下的已实现偏度RskewRet，基于量的信息构造的Rskew因子方向发生了变化，而且换手得到一个比较明显的改进。

而且相关性方面，对于FactorPool中现有的因子也有所提升和改善，而且解决了RSkewRet因子与反转因子相关性较高的一个问题。





综合来看，基于量的变化率构造的高阶矩指标是一种比较有效的改进。

**扩展二**

此处我们结合方向波动率去考虑，在第一小节中我们提到过的方向波动率这一概念，我们希望将其构造波动率的技巧运用到已实现波动（RV）的构造中

此处我们仿照波动率的定义，给出分钟频率下的方向波动率的定义

已实现上行波动率

已实现下行波动率

已实现波动率

我们进行如下测试，因子定义依次如下

|  |  |
| --- | --- |
| RVol\_020 | -ts\_Stdev(TotalRet,20) |
| RVol\_720 | -ts\_Mean(IPRetp2,20) |
| RVol\_1320 | -ts\_Mean(IPPRetp2,20) |
| RVol\_1020 | ts\_Mean(IPNRetp2,20) |

测试结果如下所示

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| RVol\_020 | 3.59% | 17.69% | 0.20 | 55.87% | 31.32% |
| RVol\_720 | 8.33% | 15.52% | 0.54 | 30.61% | 17.55% |
| RVol\_1320 | 11.44% | 15.44% | 0.74 | 26.59% | 18.32% |
| RVol\_1020 | -3.80% | 15.49% | -0.25 | 74.38% | 17.79% |

可以看出，对于已实现上行波动率和已实现下行波动率进行了分离后，这两个指标出现了分化，且剥离了下行波动率以后，上行已实现波动较已实现波动表现有所改善。

另外，我们进一步尝试了上下行波动率的组合，测试了如下因子

|  |  |
| --- | --- |
| RVol\_920 | -ts\_Stdev(IPNRetp2,20) |
| RVol\_620 | -ts\_Stdev(IPRetp2,20) |
| RVol\_1420 | -ts\_Stdev(IPPRetp2,20)./ts\_Mean(IPPRetp2,20) |
| RVol\_1520 | ts\_Mean((IPNRetp2-IPPRetp2),20) |

结果如下

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| RVol\_920 | 15.35% | 14.29% | 1.07 | 35.37% | 24.35% |
| RVol\_620 | 22.42% | 13.75% | 1.63 | 27.93% | 24.19% |
| RVol\_1420 | 30.09% | 8.75% | 3.44 | 16.95% | 37.67% |
| RVol\_1520 | 28.92% | 11.58% | 2.50 | 15.79% | 40.80% |

可以看出，单独开来表现比较平庸的两个因子，经过外部运算之后，收益表现有显著提升，远比两个因子单独作为alpha因子的表现更为优秀。

所以这种因子之间的共同构造可以说是日内已实现方向波动率构造的一种比较好构造形式。

### 日内分时段的处理

之前对于日内反转因子的研究结论显示，日内不同时间段具有不同的交易行为和交易结构，所以传统的日频反转因子无法反应日内交易行为结构的差异。

此处，我们此处具体的也有两种处理方法

**基于日内分时段数据构造日频低波动指标**

此处，我们延用日内反转因子的处理方法，使用日内特定的分时段数据，构造日频的低波动指标。

以成交额Value为例，我们将成交额分为日内4小时的成交额，Amt1h，Amt2h，Amt3h，Amt4h，成为4个独立的日频序列，对这4个独立的日频序列构造低波动因子

构造如下的低波动指标，因子定义如下：

|  |  |
| --- | --- |
| LFVol\_0Value20 | -ts\_Stdev(Value,20) |
| HFVol\_0Amt1h20 | -ts\_Stdev(Amt1h,20) |
| HFVol\_0Amt2h20 | -ts\_Stdev(Amt2h,20) |
| HFVol\_0Amt3h20 | -ts\_Stdev(Amt3h,20) |
| HFVol\_0Amt4h20 | -ts\_Stdev(Amt4h,20) |

测试结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| LFVol\_0Value20 | 36.72% | 10.01% | 3.67 | 19.39% | 23.44% |
| HFVol\_0Amt1h20 | 34.56% | 9.77% | 3.54 | 18.44% | 23.28% |
| HFVol\_0Amt2h20 | 27.77% | 9.69% | 2.87 | 20.69% | 23.02% |
| HFVol\_0Amt3h20 | 31.38% | 9.86% | 3.18 | 20.22% | 23.22% |
| HFVol\_0Amt4h20 | 39.19% | 9.85% | 3.98 | 19.29% | 23.27% |

可以看出，日内分时段数据同样呈现分化，类似地，开盘后第一小时和收盘前最后一小时构造的低波动因子会比较优秀。

**基于日内分时段数据计算波动指标**

此处我们采用日内分时段的分钟变化率数据，去构造四个日内分时段序列的标准差，例如利用第四小时的分钟收益率数据，构造日内平滑周期下的基于第四小时的收益率的标准差

因子定义如下：

|  |  |
| --- | --- |
| LFVol\_0Value20 | -ts\_Stdev(Value,20) |
| HFStd\_0StdA1h | -(StdA1h) |
| HFStd\_0StdA2h | -(StdA2h) |
| HFStd\_0StdA3h | -(StdA3h) |
| HFStd\_0StdA4h | -(StdA4h) |

测试结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| LFVol\_0Value20 | 36.72% | 10.01% | 3.67 | 19.39% | 23.44% |
| HFStd\_0StdA1h | 33.57% | 9.85% | 3.41 | 17.62% | 20.59% |
| HFStd\_0StdA2h | 24.08% | 9.98% | 2.41 | 20.87% | 19.82% |
| HFStd\_0StdA3h | 27.09% | 10.22% | 2.65 | 20.41% | 20.18% |
| HFStd\_0StdA4h | 36.70% | 10.04% | 3.65 | 19.20% | 19.57% |

可以看出，使用第二种方法，收益不如第一种方法，但是换手和波动都有所下降。

## 通用因子处理及优化

上述第二节，第三节我们讨论了几种比较有效因子的模式以及高频化的方法，他们的表现相较于日频因子有比较明显的提高。下面我们进一步细化讨论通用因子操作下对日内中心矩因子效果的影响。

### 平滑处理

该类因子用到平滑操作的处理并不多，在部分用到平滑操作的因子里，我们进行了如下测试，

因子表现如下

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| -ts\_Mean(IPLogRetp3,20) | 22.38% | 8.70% | 2.57 | 10.42% | 34.92% |
| -ts\_Decay(IPLogRetp3,20) | 25.74% | 9.01% | 2.86 | 10.87% | 37.94% |
| -ts\_DecayExp(IPLogRetp3,20) | 32.48% | 9.21% | 3.53 | 9.79% | 96.75% |

其中IPLogRetp3是日内1min对数收益率的立方和

整体来说，综合收益换手等各方面考虑，我们认为ts\_Deca作为平滑方式是一种比较有效的处理。

相比于普通一层的平滑，双重平滑的意义在于将过去里层平滑的一段时间的因子数据自我复制，然后引入外层平滑的数据新增天数的数据，其实是一种拉长平滑周期，提高内层平滑数据的权重的一种操作，适合于换手率较高，且近期数据对未来收益影响较大的因子。

而低波动因子普遍换手并不算也特别高，因此进行双重平滑并不是一个高性价比的操作。所以此处会对一些性质更接近反转因子的且平滑周期较短的类偏度因子一类因子作为双重平滑处理对象

我们以上述因子-ts\_Mean(IPLogRetp3,5)为例，

我们挑选一些双重平滑后效果比较优秀的组合，下面列出如下

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **因子名** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| -ts\_Mean(IPLogRetp3,5) | 35.28% | 9.11% | 3.87 | 9.20% | 92.46% |
| -ts\_Decay(ts\_DecayExp(IPLogRetp3,20),3) | 31.90% | 9.32% | 3.42 | 9.34% | 65.64% |
| -ts\_Decay(ts\_Decay(IPLogRetp3,20),3) | 24.38% | 9.02% | 2.70 | 10.90% | 29.78% |
| -ts\_Decay(ts\_DecayExp(IPLogRetp3,20),20) | 22.42% | 8.96% | 2.50 | 10.88% | 24.44% |

可以看出，经过处理后，虽然收益与单次平滑相比，有所下降，但是在收益损失不太严重的前提下，换手率能够有效的下降，，总体来讲还是不错的。

### 剔除常见因子的影响

过往的做法我们常用对市值或对数市值Log(股票市值)做正交化处理（横截面回归取残差），可以得到剔除了线性的市值影响的新的估值因子。使用这个新的数值，因子能选出相近市值水平中因子值较低的股票，从而避免在选股中包含市值风格影响。

但最近不少券商的研报上都会剔除主流风格因子对于因子本身的影响，其中不少有比较意思的效果

因此我们此次进行CrossFit操作的时候，将剔除因子扩展到以下几类，



测试如下

我们以-ts\_Mean((High-Low),20).\*ts\_Stdev(Volume,20)作为基础公式

观察pn\_CrossFit(-ts\_Mean((High-Low),20).\*ts\_Stdev(Volume,20),CrossData)处理下对因子的影响

测试结果如下所示

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CorssData** | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| 无 | 34.36% | 10.87% | 3.16 | 17.21% | 23.60% |
| LnCap | 20.58% | 13.23% | 1.56 | 17.35% | 20.01% |
| Mkt | 32.26% | 10.83% | 2.98 | 11.10% | 23.27% |
| BP | 34.41% | 10.42% | 3.30 | 10.12% | 23.76% |
| EP | 34.37% | 10.68% | 3.22 | 13.40% | 23.28% |
| ROE | 34.35% | 10.86% | 3.16 | 17.18% | 23.59% |
| IndRet | 33.66% | 10.10% | 3.33 | 13.60% | 62.22% |
| GrowthQ | 34.34% | 10.85% | 3.16 | 17.08% | 23.61% |
| GrowthY | 34.33% | 10.86% | 3.16 | 17.27% | 23.60% |

可以看出，大部分CrossData的效果都比较相近，能够轻微的改善sharp。但是对应各自不同的公式构造类型，结果存在差异。具体见下表



综合来说，对于中心矩因子来说，剔除EP和BP是一个比较有效的选择。

### 行业正态化

同一因子对于不同行业可能存在差异，因此可能会使构造多空组合时造成行业的偏差。进行行业正态化的处理有助于因子选出各行业中收益率最低的股票，保持组合的行业中性，去除行业风险。

具体的处理方法是，对于每一个横截面，分别找出各个行业的股票，对他们的因子值做一个正态分布转换。这种处理明确了每个股票在行业中的相对估值大小，将所有行业拉到了同一个水平线上进行比较。

而CUti中的正态转换有三种处理函数，分别是GroupNorm、GroupStd、GroupNeutral，此处都会用到

测试如下

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Value\_05 | | -ts\_Stdev(Value,5) | | | | |
| GNValue | | GroupNorm(-ts\_Stdev(Value,5),GrpLabel) | | | | |
| GNeuValue | | pn\_GroupNeutral(-ts\_Stdev(Value,5),GrpLabel) | | | | |
| GStdValue | | pn\_GroupStd(-ts\_Stdev(Value,5),GrpLabel) | | | | |
|  | **年化收益** | | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** | |
| Value\_05 | 37.07% | | 10.75% | 3.45 | 17.47% | 71.03% | |
| GNValue | 35.34% | | 8.24% | 4.29 | 7.89% | 75.58% | |
| GNeuValue | 32.29% | | 7.66% | 4.21 | 6.04% | 66.62% | |
| GStdValue | 34.78% | | 8.01% | 4.34 | 7.85% | 73.48% | |

回测比较符合预期，整体来说，行业操作能降低波动，提高sharp，在结合适收益与Sharp保留来看

整体来讲，对于波动率因子来说，GroupNorm>GroupStd>GroupNeutral，但通常三者差异并不是太大。

特别的，对于基于长短周期的波动率因子来说，GroupNorm失效

### 极端值去除

部分因子的原始数据可能存在极端值数据不可信、极端值不可用模型估量等现象，或者极端估值是由非市场波动因素造成因此不会平复。由于根据因子值构建的多空组合要求因子值越大收益越高，去除这些无序的极端值部分有助于提高因子多空组合的收益。

测试如下

HLVolume\_0：不截尾

HLVolume\_101：截尾1%

HLVolume\_105：截尾5%

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **年化收益** | **年化波动** | **夏普比率** | **最大回撤** | **日均换手** |
| HLVolume\_0 | 34.36% | 10.87% | 3.16 | 17.21% | 23.60% |
| HLVolume\_101 | 33.02% | 10.23% | 3.23 | 14.50% | 24.84% |
| HLVolume\_105 | 29.40% | 8.79% | 3.35 | 9.54% | 28.98% |

可以看出，该类因子不适合截尾操作

## 批量测试

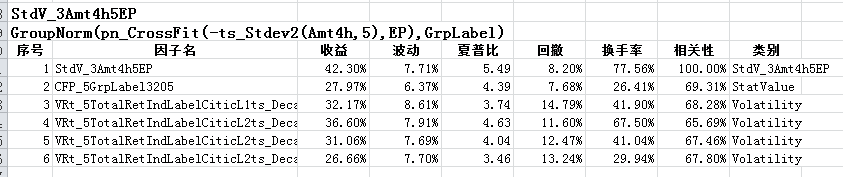
具体的批量测试采取综合上述第四节一些有效因子构造和有效的因子处理，进行批量回测和参数寻优，并利用GetForntier筛选出优超的指标。具体结果由Excel输出，并与FactorPool的相关性进行比较。

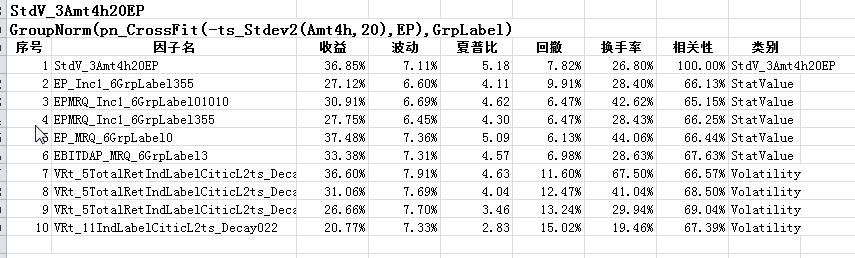
HFStd\_v7\_5是经过筛选后的因子测试表现，目前根据已有的回测结果来看，目前做出来的中心矩因子，存在部分因子是FacotPool中没有覆盖的，或者可以对FactorPool中部分已有因子形成优超的。

下面挑选一些回测后一些比较有价值的因子列出如下，详细相关结果以及与FactorPool的比较参见Excel。

**收益和Sharp较高**

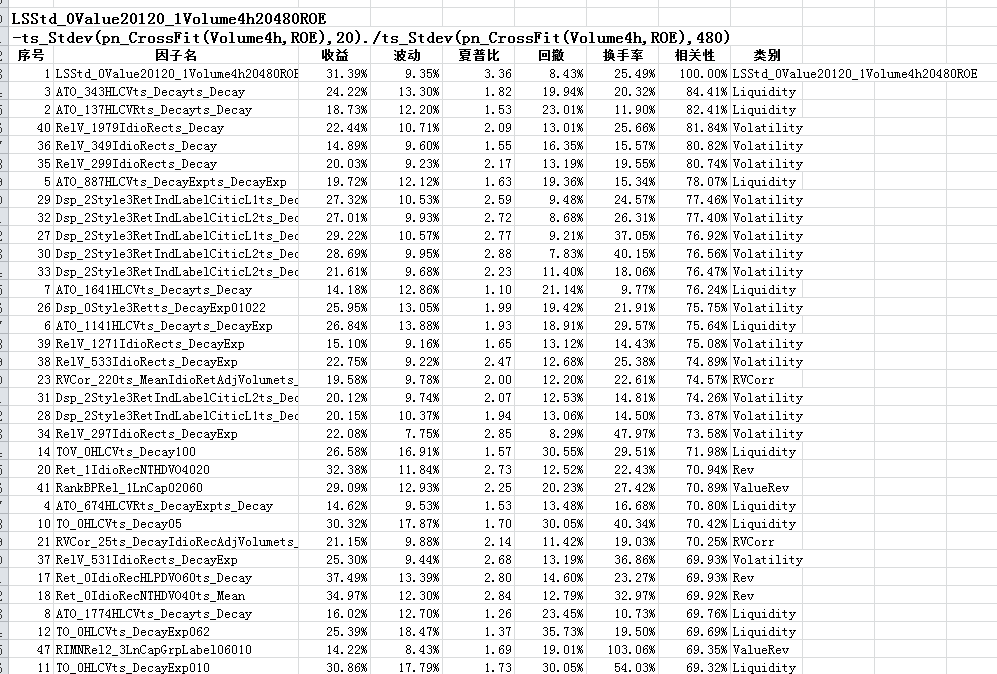
**GroupNorm(pn\_CrossFit(-ts\_Stdev2(Amt4h,5),EP),GrpLabel)**

 **GroupNorm(pn\_CrossFit(-ts\_Stdev2(Amt4h,20),EP),GrpLabel)**

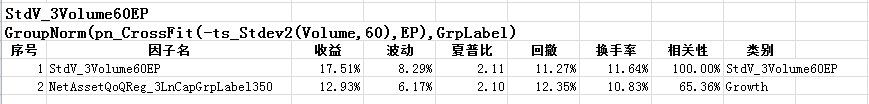


**收益换手比高**

**-ts\_Stdev(pn\_CrossFit(Volume4h,ROE),20)./ts\_Stdev(pn\_CrossFit(Volume4h,ROE),480)**



**GroupNorm(pn\_CrossFit(-ts\_Stdev2(Volume,60),EP),GrpLabel)**

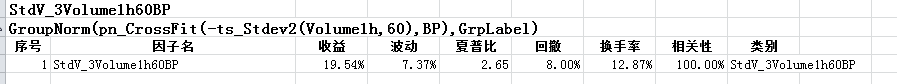


**pn\_CrossFit(-ts\_Stdev2(Volume1h,20),BP)**

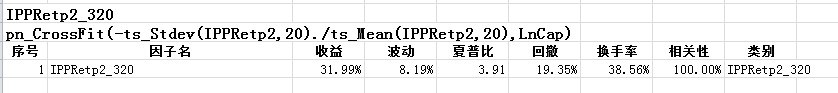


**与FactorPool现有因子相关性较低**

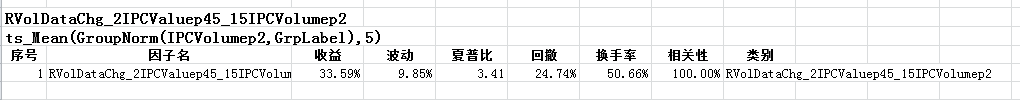
**GroupNorm(pn\_CrossFit(-ts\_Stdev2(Volume1h,60),BP),GrpLabel)**



**pn\_CrossFit(-ts\_Stdev(IPPRetp2,20)./ts\_Mean(IPPRetp2,20),LnCap)**



**ts\_Mean(GroupNorm(IPCVolumep2,GrpLabel),5)**



综合回测来看，中心矩因子中的低波动因子是一种比较稳定的因子，多空收益不如反转因子优秀，但是换手比较温和。使用日内分时段数据结合行业操作和CrossFit操作之后能得到比较稳健的因子。

而日内的已实现偏度因子的性质上会更接近反转因子，但是波动会更低一点，利用日内分钟量价数据的变化率构造的已实现因子能够带来目前FactorPool没有覆盖的新的一种因子。

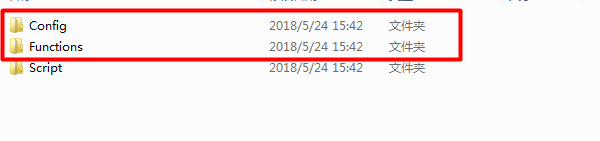
由于此次测试参数比较多，囿于测试时间和效率的限制，所以部分公式参数无法遍历完成，只能小规模回测参数寻优后确定一个大致的最优范围。所以后续研究可以进一步精细化寻优。

## 函数说明

本次需要配置的函数主要是三部分，分列如下

1. 高频计算函数（按以山哥修改后为基础的0524提交的最新版作为基准）

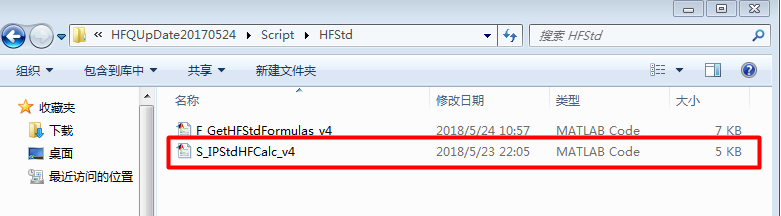
此部分函数属于通常配置，以后计算其他日内因子也需要调用该系列函数，主要包括下图红框的两个文件夹中的函数，是仿照日频回测中的CUti和F\_ProduceValue几个因子产生函数改造而成的高频版本。



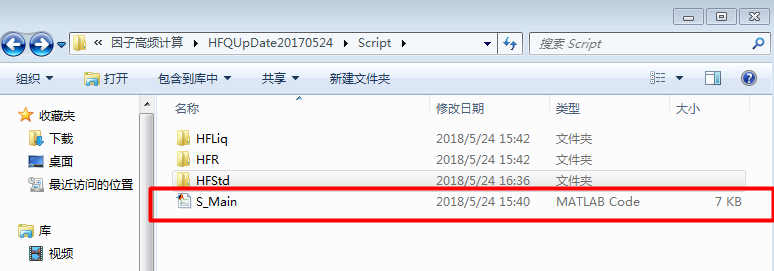
1. 日内波动因子高频计算函数

此处是调用HFQ的Bar数据，调用模式有两个

1. 可以调用S\_IPStdHFCalc\_v4单独生成日内波动率因子的中间变量

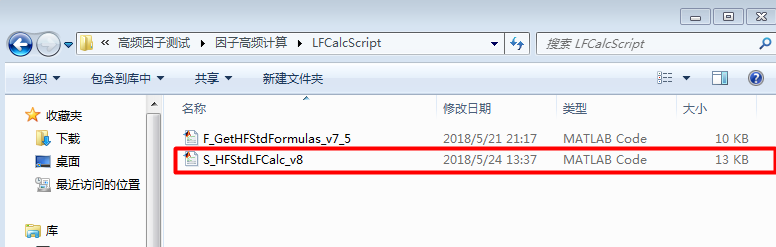


1. 可以直接调用S\_Main直接生成



1. 日内波动性因子低频回测函数

此部分函数是调用2中生成中间文件，进行进一步的加工处理得到最后的因子值，可以在配置了日频回测平台上的计算机中运行。主函数是S\_HFStdLFCalc\_v8



## 参考文献

《华泰单因子系列之五单因子测试之波动率因子》

《国君数量化专题一零八—基于不同域研究的多因子选股体系》

《东方证券-金融工程研究技术类新Alpha因子的批量测试 20170217》

《西南证券-基于方向波动率的选股因子研究20171221》

《国泰君安数量化专题之九十三：基于短周期价量特征的多因子选股体系》

《基于加权已实现极差的中国股市波动特征\_文凤华》

《加权已实现极差四次幂变差分析及其应用\_唐勇》

《中国股票市场日内波动率研究\_王国华》

《引入隔夜信息的已实现波动率\_瞿慧》

## 后续课题的列表

根据昨天跟山哥讨论后的结果，整理了正在进行（待完成）以及后续可能拓展研究的工作方向，列出如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **课题名称** | **研究基础** | **数据限制** | **研究类型** | **课题不确定性** |
| 反转因子与宏观因子的相关性分析初探 | 进行中 | 无 | 因子方法论改造  择时研究 | 中等 |
| 日内量价形态因子RVCorr的研究 | 有 | 无 | 因子研究 | 偏低 |
| 日内群组动量因子MoM的研究 | 有 | 无 | 因子研究 | 偏低 |
| 高频Hurst因子研究 | 无 | 无 | 因子研究 | 高 |
| 个股筹码分布因子与区间突破的研究 | 无 | 无 | 因子研究 | 中等 |
| 个股独立行情因子的研究 | 前期研究 | 无 | 因子研究 | 偏低 |
| 盘口挂单隐藏流动性因子HIddenLiq的研究 | 无 | 部分限制 | 因子研究  日内策略的储备 | 高 |
| 日内病态流动性因子（聪明钱）的研究 | 前期研究 | 无 | 因子研究 | 偏低 |
| 基于买卖结构的现金流因子的再研究 | 前期研究 | 有限制 | 因子研究 | 偏低 |
| 因子方法改进 | 无 | 无 | 因子方法论改造 | 高 |

对于上述课题的具体研究内容，进行如下概述

1. 反转因子与宏观因子的相关性分析初探

整理宏观基本面数据，并与反转因子的阶段表现进行相关性分析。找寻并整理与反转因子相关性较高的基本面指标，尝试将其中稳定性较高的指标加入现有选股预测体系中或加入到因子构造中，通过批量回测，可以找寻出一些有效的利用基本面指标对市场状态进行区分的方法

1. 日内量价形态因子RVCorr的研究
2. 利用研报中或FactorPool中现有的量价相关指标，进行高频化改造，构造日内RVCorr指标
3. 整理总结归纳日内形态特征，并思考相应特征的量化处理方法
4. 日内群组动量MOM因子的研究

基于其他研究员前期的研究，将MOM因子进行高频化改造

1. 高频Hurst因子的研究

Hurst指数是描述时间序列是否服从随机游走的重要工具，能够刻画时间序列的长记忆性和识别长期趋势，可能是一个重要的对个股进行区分的分层指标。目前大多卖方研究Hurst指数基于日频序列，计算过程样本太小，学术上使用日内数据可以大大提高该指标的准确性。

1. 个股筹码集中度分布和区间突破

通过刻画股票的买入成本来构造Alpha因子

1. 个股独立行情因子

关注个股与行业其他个股日内表现背离的股票，并以此构造Alpha因子

1. 盘口挂单隐藏流动性的研究

利用盘口隐藏挂单对潜在买卖力量构建Alpha指标，严格意义上应该使用Tick数据。但是可以利用Bar数据进行近似处理。

1. 日内病态流动性因子的研究

通过单位资金撬动的收益这一病态流动ILLiq指标，用以区分市场上的激进交易和普通交易，并以此为构造两种不同交易下的同一指标的相对位置Alpha因子

1. 基于买卖结构的现金流因子的再研究

根据主卖主买指标，将该指标融入到已有的日内因子构造中，能够进一步提升收益。

1. 因子方法论研究

目前研究的高频因子与反转因子有比较高的相关性，而反转因子的问题主要是如下来那个方面

其一，多头部分收益不如空头部分

其二，因子在某些行业表现较差，诸如银行和非银

目前有几个可以探索的改进方向

其一，因子的分层效应与分层处理，可能引入分层因子对反转指标的某一分位数段位进行分层处理。

其二，相关性较高指标的互补，包括可以考虑同一指标不同参数在某一表现较佳行业中对现有指标的替代。以及负相关较高的指标在该行业的互补构造。

其三，行业内部选股逻辑的再优化，例如银行和非银，是有独特的专属指标，结合这些指标可以重新构造因子在这些行业的独特逻辑。对行业内选股建模进行更细化的考量。